Package 'uclust'

January 20, 2020

Title Clustering and Classification Inference with U-Statistics

Version 0.2.0

Description Clustering and classification inference for high dimension low sample size (HDLSS) data with U-statistics. The package contains implementations of nonparametric statistical tests for sample homogeneity, group separation, clustering, and classification of multivariate data. The methods have high statistical power and are tailored for data in which the dimension L is much larger than sample size n. See Gabriela B. Cybis, Marcio Valk and Sílvia RC Lopes (2018) <doi:10.1080/00949655.2017.1374387> and Marcio Valk and Gabriela B. Cybis (2018) <arXiv:1805.12179>.

Depends R (>= 3.4.0),dendextend,robcor

Imports

License GPL-3

Encoding UTF-8

LazyData true

RoxygenNote 7.0.1

Suggests testthat

NeedsCompilation no

Author Gabriela Cybis [aut, cre],
Marcio Valk [aut],
Kazuki Yokoyama [ctb]

Maintainer Gabriela Cybis <gcybis@gmail.com>

Repository CRAN

Date/Publication 2020-01-20 18:30:02 UTC

R topics documented:

---

bn *Computes Bn Statistic.*

Description

Returns the value for the Bn statistic that measures the degree of separation between two groups. The statistic is computed through the difference of average within group distances to average between group distances. Large values of Bn indicate large group separation. Under overall sample homogeneity we have E(Bn)=0.

Usage

```
bn(group_id, md = NULL, data = NULL)
```

Arguments

`group_id`	A vector of 0s and 1s indicating to which group the samples belong. Must be in the same order as data or md.
`md`	Matrix of distances between all data points.
`data`	Data matrix. Each row represents an observation.

Details

Either `data` OR `md` should be provided. If data are entered directly, Bn will be computed considering the squared Euclidean distance, which is compatible with `is_homo`, `uclust` and `uhclust`.

For more detail see Cybis, Gabriela B., Marcio Valk, and Sílvia RC Lopes. "Clustering and classification problems in genetics through U-statistics." Journal of Statistical Computation and Simulation 88.10 (2018) and Valk, Marcio, and Gabriela Bettella Cybis. "U-statistical inference for hierarchical clustering." arXiv preprint arXiv:1805.12179 (2018).

Value

Value of the Bn statistic.

Examples

```
n=5
x=matrix(rnorm(n*10),ncol=10)
bn(c(1,0,0,0,0),data=x)     # option (a) entering the data matrix directly
md=as.matrix(dist(x))^2
bn(c(0,1,1,1,1),md)         # option (b) entering the distance matrix
```

`is_homo` *U-statistic based homogeneity test*

Description

Homogeneity test based on the statistic `bn`. The test assesses whether there exists a data partition for which group separation is statistically significant according to the U-test. The null hypothesis is overall sample homogeneity, and a sample is considered homogeneous if it cannot be divided into two statistically significant subgroups.

Usage

```
is_homo(md = NULL, data = NULL, rep = 10)
```

Arguments

`md`	Matrix of squared Euclidean distances between all data points.
`data`	Data matrix. Each row represents an observation.
`rep`	Number of times to repeat optimization procedure. Important for problems with multiple optima.

Details

This is the homogeneity test of Cybis et al. (2017) extended to account for groups of size 1. The test is performed through two steps: an optimization procedure that finds the data partition that maximizes the standardized Bn and a test for the resulting maximal partition. Should be used in high dimension small sample size settings.

Either `data` or `md` should be provided. If data are entered directly, Bn will be computed considering the squared Euclidean distance. It is important that if a distance matrix is entered, it consists of squared Euclidean distances, otherwise test results are invalid.

Variance of `bn` is estimated through resampling, and thus, p-values may vary a bit in different runs.

For more detail see Cybis, Gabriela B., Marcio Valk, and Sílvia RC Lopes. "Clustering and classification problems in genetics through U-statistics." Journal of Statistical Computation and Simulation 88.10 (2018) and Valk, Marcio, and Gabriela Bettella Cybis. "U-statistical inference for hierarchical clustering." arXiv preprint arXiv:1805.12179 (2018).

Value

Returns a list with the following elements:

minFobj Test statistic. Minimum of the objective function for optimization (-stdBn).

group1 Elements in group 1 in the maximal partition. (obs: this is not the best partition for the data, see `uclust`)

group2 Elements in group 2 in the maximal partition.

p.MaxTest P-value for the homogeneity test.

Rep.Fobj Values for the minimum objective function on all `rep` optimization runs.

bootB Resampling variance estimate for partitions with groups of size n/2 (or (n-1)/2 and (n+1)/2 if n is odd).

bootB1 Resampling variance estimate for partitions with one group of size 1.

Examples

```
x = matrix(rnorm(500000),nrow=50)  #creating homogeneous Gaussian dataset
res = is_homo(data=x)

x[1:30,] = x[1:30,]+0.15   #Heterogeneous dataset (first 30 samples have different mean)
res = is_homo(data=x)

md = as.matrix(dist(x)^2)   #squared Euclidean distances for the same data
res = is_homo(md)

# Multidimensional sacling plot of distance matrix
fit <- cmdscale(md, eig = TRUE, k = 2)
x <- fit$points[, 1]
y <- fit$points[, 2]
plot(x,y, main=paste("Homogeneity test: p-value =",res$p.MaxTest))
```

`plot_uhclust` *Plot function for the result of uhclust*

Description

This function plots the p-value annotated dendrogram resulting from `uhclust`

Usage

```
plot_uhclust(
  uhclust,
  pvalues_cex = 0.8,
  pvalues_dx = 2,
  pvalues_dy = 0.08,
  print_pvalues = TRUE
)
```

Arguments

uhclust	Result from uhclust
pvalues_cex	Graphical parameter for p-value font size.
pvalues_dx	Graphical parameter for p-value position shift on x axis.
pvalues_dy	Graphical parameter for p-value position shift on y axis.
print_pvalues	Logical. Should the p-values be printed?

Examples

```
x = matrix(rnorm(100000),nrow=50)
x[1:35,] = x[1:35,]+0.7
x[1:15,] = x[1:15,]+0.4
res = uhclust(data=x, plot=FALSE)
plot_uhclust(res)
```

---

print.utest_classify *Simple print method for utest_classify objects.*

Description

Simple print method for utest_classify objects.

Usage

```
## S3 method for class 'utest_classify'
print(x, ...)
```

Arguments

x	utest_classify object
...	additional parameters passed to the function

rep_optimBn *Optimization function with multiple staring points (for local optima)*

Description

Finds the configuration with max Bn among all configurations.

Usage

```
rep_optimBn(mdm, rep = 15, bootB = -1)
```

Arguments

mdm	Matrix of squared Euclidean distances between all data points.
rep	Number of replications
bootB	Result of previous bootstrap (if available). If, -1, a new bootstrap is performed for the variance of Bn.

uclust *U-statistic based significance clustering*

Description

Partitions the sample into the two significant subgroups with the largest Bn statistic. If no significant partition exists, the test will return "homogeneous".

Usage

```
uclust(md = NULL, data = NULL, alpha = 0.05, rep = 15)
```

Arguments

md	Matrix of squared Euclidean distances between all data points.
data	Data matrix. Each row represents an observation.
alpha	Significance level.
rep	Number of times to repeat optimization procedures. Important for problems with multiple optima.

Details

This is the significance clustering procedure of Valk and Cybis (2018). The method first performs a homogeneity test to verify whether the data can be significantly partitioned. If the hypothesis of homogeneity is rejected, then the method will search, among all the significant partitions, for the partition that better separates the data, as measured by larger bn statistic. This function should be used in high dimension small sample size settings.

Either `data` or `md` should be provided. If data are entered directly, Bn will be computed considering the squared Euclidean distance. It is important that if a distance matrix is entered, it consists of squared Euclidean distances, otherwise test results are invalid.

Variance of bn is estimated through resampling, and thus, p-values may vary a bit in different runs.

For more detail see Cybis, Gabriela B., Marcio Valk, and Sílvia RC Lopes. "Clustering and classification problems in genetics through U-statistics." Journal of Statistical Computation and Simulation 88.10 (2018) and Valk, Marcio, and Gabriela Bettella Cybis. "U-statistical inference for hierarchical clustering." arXiv preprint arXiv:1805.12179 (2018). See also `is_homo`, `uhclust`, `Utest_class`.

Value

Returns a list with the following elements:

cluster1 Elements in group 1 in the final partition. This is the significant partition with maximal Bn, if sample is heterogeneous.

cluster2 Elements in group 2 in the final partition.

p.value P-value for the test that renders the final partition, if heterogeneous. Homogeneity test p-value, if homogeneous.

alpha_corrected Bonferroni corrected significance level for the test that renders the final partition, if heterogeneous. Homogeneity test significance level, if homogeneous.

n1 Size of the smallest cluster

ishomo Logical, returns `TRUE` when the sample is homogeneous.

Bn Value of Bn statistic for the final partition, if heterogeneous. Value of Bn statistic for the maximal homogeneity test partition, if homogeneous.

varBn Variance estimate for final partition, if heterogeneous. Variance estimate for the maximal homogeneity test partition, if homogeneous.

ishomoResult Result of homogeneity test (see `is_homo`).

Examples

```
set.seed(17161)
x = matrix(rnorm(100000),nrow=50)  #creating homogeneous Gaussian dataset
res = uclust(data=x)

x[1:30,] = x[1:30,]+0.25   #Heterogeneous dataset (first 30 samples have different mean)
res = uclust(data=x)

md = as.matrix(dist(x)^2)   #squared Euclidean distances for the same data
res = uclust(md)
```

```
# Multidimensional scaling plot of distance matrix
fit <- cmdscale(md, eig = TRUE, k = 2)
x <- fit$points[, 1]
y <- fit$points[, 2]
col=rep(3,dim(md)[1])
col[res$cluster2]=2
plot(x,y, main=paste("Multidimensional scaling plot of data:
                    homogeneity p-value =",res$ishomoResult$p.MaxTest),col=col)
```

---

uhclust *U-statistic based significance hierarchical clustering*

Description

Hierarchical clustering method that partitions the data only when these partitions are statistically significant.

Usage

```
uhclust(md = NULL, data = NULL, alpha = 0.05, rep = 15, plot = TRUE)
```

Arguments

md	Matrix of squared Euclidean distances between all data points.
data	Data matrix. Each row represents an observation.
alpha	Significance level.
rep	Number of times to repeat optimization procedures. Important for problems with multiple optima.
plot	Logical, TRUE if p-value annotated dendrogram should be plotted.

Details

This is the significance hierarchical clustering procedure of Valk and Cybis (2018). The data are repeatedly partitioned into two subgroups, through function uclust, according to a hierarchical scheme. The procedure stops when resulting subgroups are homogeneous or have fewer than 3 elements. This function should be used in high dimension small sample size settings.

Either data or md should be provided. If data are entered directly, Bn will be computed considering the squared Euclidean distance. It is important that if a distance matrix is entered, it consists of squared Euclidean distances, otherwise test results are invalid.

Variance of bn is estimated through resampling, and thus, p-values may vary a bit in different runs.

For more detail see Cybis, Gabriela B., Marcio Valk, and Sílvia RC Lopes. "Clustering and classification problems in genetics through U-statistics." Journal of Statistical Computation and Simulation 88.10 (2018) and Valk, Marcio, and Gabriela Bettella Cybis. "U-statistical inference for hierarchical clustering." arXiv preprint arXiv:1805.12179 (2018).

See also is_homo, uclust and Utest_class.

Value

Returns an object of class `hclust` with three additional attribute arrays:

Pvalues P-values from uclust for the final data partition at each node of the dendrogram. This array is in the same order of `height`, and only contains values for tests that were performed.

alpha Bonferroni corrected significance levels for uclust for the data partitions at each node of the dendrogram. This array is in the same order of `height`, and only contains values for tests that were performed.

groups Final group assignments.

Examples

```
x = matrix(rnorm(100000),nrow=50)  #creating homogeneous Gaussian dataset
res = uhclust(data=x)


x[1:30,] = x[1:30,]+0.7   #Heterogeneous dataset
x[1:10,] = x[1:10,]+0.4
res = uhclust(data=x)
res$groups
```

---

utest *U test*

Description

Test for the separation of two groups. The null hypothesis states that the groups are homogeneous and the alternative hypothesis states that they are separate.

Usage

```
utest(group_id, md = NULL, data = NULL, numB = 1000)
```

Arguments

`group_id`	A vector of 0s and 1s indicating to which group the samples belong. Must be in the same order as data or md.
`md`	Matrix of distances between all data points.
`data`	Data matrix. Each row represents an observation.
`numB`	Number of resampling iterations.

Details

Either `data` or `md` should be provided. If data are entered directly, Bn will be computed considering the squared Euclidean distance, which is compatible with `is_homo`, `uclust` and `uhclust`.

For more details see Cybis, Gabriela B., Marcio Valk, and Sílvia RC Lopes. "Clustering and classification problems in genetics through U-statistics." Journal of Statistical Computation and Simulation 88.10 (2018)

Value

Returns a list with the following elements:

Bn Test Statistic

Pvalue Replication based p-value

Replication Number of replications used to compute p-value

See Also

`bn`,`is_homo`

Examples

```
# Simulate a dataset with two separate groups, the first 5 rows have mean 0 and
# the last 5 rows have mean 5.
data <- matrix(c(rnorm(75, 0), rnorm(75, 5)), nrow = 10, byrow=TRUE)

# U test for mixed up groups
utest(group_id=c(1,0,1,0,1,0,1,0,1,0), data=data, numB=3000)
# U test for correct group definitions
utest(group_id=c(1,1,1,1,1,0,0,0,0,0), data=data, numB=3000)
```

---

utest_classify *Test for classification of a sample in one of two groups.*

Description

The null hypothesis is that the new data is not well classified into the first group when compared to the second group. The alternative hypothesis is that the data is well classified into the first group.

Usage

```
utest_classify(x, data, group_id, bootstrap_iter = 1000)
```

Arguments

x	A numeric vector to be classified.
data	Data matrix. Each row represents an observation.
group_id	A vector of 0s (first group) and 1s indicating to which group the samples belong. Must be in the same order as data.
bootstrap_iter	Numeric scalar. The number of bootstraps. It's recommended $1000 < bootstrap_iter < 10000$.

Details

The test is performed considering the squared Euclidean distance.

For more detail see Cybis, Gabriela B., Marcio Valk, and Sílvia RC Lopes. "Clustering and classification problems in genetics through U-statistics." Journal of Statistical Computation and Simulation 88.10 (2018) and Valk, Marcio, and Gabriela Bettella Cybis. "U-statistical inference for hierarchical clustering." arXiv preprint arXiv:1805.12179 (2018).

Value

A list with class "utest_classify" containing the following components:

statistic	the value of the test statistic.
p_value	The p-value for the test.
bootstrap_iter	the number of bootstrap iterations.

Examples

```
# Example 1
# Five observations from each group, G1 and G2. Each observation has 60 dimensions.
data <- matrix(c(rnorm(300, 0), rnorm(300, 10)), ncol = 60, byrow=TRUE)
# Test data comes from G1.
x <- rnorm(60, 0)
# The test correctly indicates that the test data should be classified into G1 (p < 0.05).
utest_classify(x, data, group_id = c(rep(0,times=5),rep(1,times=5)))

# Example 2
# Five observations from each group, G1 and G2. Each observation has 60 dimensions.
data <- matrix(c(rnorm(300, 0), rnorm(300, 10)), ncol = 60, byrow=TRUE)
# Test data comes from G2.
x <- rnorm(60, 10)
# The test correctly indicates that the test data should be classified into G2 (p > 0.05).
utest_classify(x, data, group_id = c(rep(1,times=5),rep(0,times=5)))
```

var_bn *Variance of Bn*

Description

Estimates the variance of the Bn statistic using the resampling procedure described in Cybis, Gabriela B., Marcio Valk, and Sílvia RC Lopes. "Clustering and classification problems in genetics through U-statistics." Journal of Statistical Computation and Simulation 88.10 (2018) and Valk, Marcio, and Gabriela Bettella Cybis. "U-statistical inference for hierarchical clustering." arXiv preprint arXiv:1805.12179 (2018).

Usage

```
var_bn(group_sizes, md = NULL, data = NULL, numB = 2000)
```

Arguments

group_sizes	A vector with two entries: size of group 1 and size of group 2.
md	Matrix of distances between all data points.
data	Data matrix. Each row represents an observation.
numB	Number of resampling iterations. Only used if no groups are of size 1.

Details

Either data or md should be provided. If data are entered directly, Bn will be computed considering the squared Euclidean distance, which is compatible with is_homo, uclust and uhclust.

Value

Variance of Bn

See Also

bn

Examples

```
n=5
x=matrix(rnorm(n*20),ncol=20)
# option (a) entering the data matrix directly and considering a group of size 1
var_bn(c(1,4),data=x)

# option (b) entering the distance matrix and considering a groups of size 2 and 3
md=as.matrix(dist(x))^2
var_bn(c(2,3),md)
```

Index